Sur le principe d'optimalité de Bellman pour les zs-POSG

Olivier Buffet, Jilles Dibangoye, Aurélien Delage, Abdallah Saffidine, Vincent Thomas

Sur le principe d'optimalité de Bellman pour les zs-POSG

Olivier Buffet[1] Jilles Dibangoye[2] Aurélien Delage[1,2] Abdallah Saffidine[3] Vincent Thomas[1]

[1] Université de Lorraine, CNRS, INRIA, LORIA, F-54000 Nancy
[2] Université de Lyon, INSA Lyon and Inria, CITI, F-69000 Lyon
[3] University of New South Wales, Sydney, Australie

(prenom.nom@inria.fr|abdallahs[@]cse.unsw.edu.au)

Résumé

*De nombreux problèmes de prise de décision séquentielle sont résolus efficacement en exploitant le principe d'optimalité de Bellman, c'est-à-dire l'imbrication récursive de sous-problèmes dans le problème original. Nous montrons ici qu'il peut être appliqué aux jeux stochastiques partiellement observables à 2 joueurs et somme nulle (zs-POSG) en (i) prenant le point de vue d'un planificateur central, qui ne peut raisonner que sur une statistique suffisante appelée état d'occupation, et (ii) transformant de tels problèmes en des jeux de Markov dans l'espace des «états d'occupation» à somme nulle (zs-OMG). Ensuite, en exploitant des propriétés de Lipschitz-continuité de la fonction de valeur optimale, on peut dériver une version de l'algorithme HSVI (*Heuristic Search Value Iteration*) qui trouve un ϵ-équilibre de Nash en temps fini.*

Mots Clef

POSG ; jeux stochastiques partiellement observables ; principe d'optimalité de Bellman ; Heuristic Search Value Iteration.

Abstract

Many non-trivial sequential decision-making problems are efficiently solved by exploiting Bellman's optimality principle, i.e.*, the recursive nesting of sub-problems within the original problem. Here we show how it can apply to (infinite horizon) 2-player zero-sum partially observable stochastic games (zs-POSGs) by (i) taking a central planner's viewpoint, which can only reason on a sufficient statistic called occupancy state, and (ii) turning such problems into zero-sum occupancy Markov games (zs-OMGs). Then, exploiting Lipschitz-continuity properties of the optimal value function, one can derive a version of the HSVI algorithm (Heuristic Search Value Iteration) that provably finds an ϵ-Nash equilibrium in finite time.*

Keywords

POSG ; partially observable stochastic games ; Bellman's optimality principle ; Heuristic Search Value Iteration.

[Note : Cet article résume un document à paraître [6].]

1 Introduction

Le principe d'optimalité de Bellman [3] a conduit à des solveurs de référence pour de nombreux problèmes de prise de décision séquentielle non-triviaux, supposant une observabilité partielle [19], des critères multi-objectifs [22, 16], des agents collaborant, par exemple modélisés par des processus de décision markoviens partiellement observables décentralisés (Dec-POMDP) [12, 23, 10], ou certains jeux non-collaboratifs (à commencer par le travail précurseur de Shapley [20], voir aussi [7]). Dans chacun de ces cadres ce principe exploite l'imbrication récursive de sous-problèmes dans le problème original. La question reste ouverte de savoir si — et comment — ce principe pourrait être appliqué aux jeux à information imparfaite, lesquels sont rencontrés dans des applications diverses telles que le Poker [15] ou les jeux de sécurité [1]. Ce travaille répond à cette question dans le cadre des jeux stochastiques partiellement observables à 2-joueurs et somme nulle (zs-POSG), *c.-à-d.* des jeux à information imparfaite, actions simultanées, mémoire parfaite (*perfect recall*), récompenses atténuées et horizon temporel potentiellement infini.

2 État de l'art

Comme les POSG et Dec-POMDP généraux, les zs-POSG à horizon infini sont indécidables, et leurs approximations à horizon fini sont dans NEXP [17, 4]. Les techniques de résolution pour POSG à horizon fini, ou autres jeux à information imparfaite qui peuvent être formulés comme des jeux en forme extensive (EFG), résolvent typiquement un programme linéaire [21], par exemple dérivé du jeu en forme normale équivalent, ou emploient un mécanisme de minimisation du regret dédié [25, 5]. Ils ne reposent donc pas sur le principe d'optimalité de Bellman, sauf (i) une approche par programmation dynamique qui ne fait que construire des ensembles de solutions non dominées [12], (ii) dans les problèmes collaboratifs (POMDP décentralisés), en adoptant le point de vue d'un planificateur central [23, 10], et (iii) pour des cadres (majoritairement à 2 joueurs et somme nulle) sous des hypothèses d'observabilité telles que l'on peut raisonner sur les croyances des joueurs [11, 8, 2, 14, 9, 13]. Ici, nous ne faisons pas d'hypothèses autres que le jeu étant

à 2 joueurs et somme nulle, en particulier concernant l'observabilité de l'état et des actions.

3 Des zs-POSG aux zs-OMG

Comme dans nombre de solveurs de Dec-POMDP, notre approche adopte le point de vue non d'un joueur, mais d'un planificateur central (hors-ligne) qui préscrit une stratégie individuelle à chaque joueur [23] en raisonnant sur les profils de stratégies partielles exécutés de $t = 0$ à τ. Le planificateur fait ainsi face à un jeu dont la dynamique est déterministe, puisqu'une stratégie partielle évolue de manière déterministe quand on la complète d'une paire de règles de décision sur un pas de temps, et dans lequel les actions des joueurs (prises non pas dans les ensembles d'actions de départ, mais parmi les règles de décision choisies par l'un et l'autre joueur) sont publiques. Ainsi, le profil de stratégies partielles courant est toujours connu des deux joueurs.

Validité du principe d'optimalité de Bellman Une statistique introduite initialement pour résoudre les Dec-POMDP, l'*état d'occupation* o_τ [10] — *c.-à-d.* la distribution de probabilité sur (i) l'état du système et (ii) l'historique d'action-observation jointe des deux joueurs étant donnée un profil de stratégies partielles — est alors employée pour démontrer que

1. étant donné un profil de stratégies partielles, on peut trouver, à l'aide de l'état d'occupation correspondant, un profil de stratégies optimal des pas de temps τ à H (l'horizon temporel fini) ;
2. toute solution optimale de ce jeu est construite récursivement sur des solutions optimales de sous-problèmes imbriqués, ce qui correspond au principe d'optimalité de Bellman ; et
3. l'état d'occupation est une statistique suffisante (en remplacement du profil de stratégies partielles) pour résoudre de manière optimale le jeu auquel est confronté le planificateur central.

Dans la suite, nous dénotons $V^*_\tau(o_\tau)$ la valeur optimale (à l'équilibre de Nash) pour le sous-problème rencontré dans l'état d'occupation o_τ, et appelons $V^*_\tau(\cdot)$ la *fonction de valeur optimale*.

Des jeux en forme anormale Le jeu en question est appelé un *occupancy Markov Game* (OMG). [1] Un OMG n'est toutefois pas un jeu de Markov standard, non seulement parce que l'espace d'états est continu et la dynamique déterministe, mais aussi parce que les *jeux locaux* dans les états d'occupation rencontrés, où l'on optimise le profil de règles de décision immédiates $\boldsymbol{\beta}_\tau$, ne sont pas des jeux en forme normale, de sorte qu'ils ne peuvent être résolus via un LP comme pour les jeux en forme normale à 2 joueurs et somme nulle. On peut toutefois raisonner sur des *sous-jeux*, dans lesquels on optimise non plus les seules règles de décision immédiate, mais toutes jusqu'à $H-1$, sous-jeux qui sont, eux, équivalents à des jeux en forme normale. Cela permet de prouver que les valeurs maximin et minimax des jeux locaux en forme **a**normale sont égales, correspondant donc à la valeur unique des équilibres de Nash, et induisent un profil de stratégies en équilibre de Nash.

La section suivante introduit des ingrédients clefs qui rendent possible la résolution de zs-OMG en exploitant le principe d'optimalité de Bellman.

4 Exploiter la structure de V^* et Q^*

Deux difficultés qui nous empêchent pour l'instant de résoudre des zs-OMG sont que :

1. à cause des espaces continus d'états et d'actions, on ne peut résoudre le problème en explorant l'infiniment-ramifié (*infinitely-branching*) arbre des futurs possibles sans capacités de généralisation ; et
2. il nous manque une solution pour résoudre un jeu en forme anormale donné.

Dans la suite, nous exploitons les propriétés des fonctions de valeur optimales V^* et Q^* pour traiter ces deux difficultés.

Lipschitz-continuité de V^* et Q^* Les propriétés de linéarité et de Lipschitz-continuité (LC) des fonctions de transition et de récompense du zs-OMG permettent de démontrer que, dans le cas d'un horizon temporel fini,

— la fonction de valeur optimale $V^*_\tau(o_\tau)$ est LC dans l'espace des états d'occupation ; et

— la fonction d'action-valeur optimale $Q^*_\tau(o_\tau, \beta^1_\tau, \beta^2_\tau)$ est LC dans les espaces des états d'occupation o_τ comme des règles de décision individuelles β^1_τ et β^2_τ.

Résolution de jeux en forme *a*normale $Q^*_\tau(o_\tau, \cdot, \cdot)$ définissant un jeu en forme anormale à résoudre en o_τ, ses propriétés de Lipschitz-continuité permettent d'envisager une procédure d'optimisation bi-niveau reposant sur un algorithme tel que DOO (*Deterministic Optimistic Optimization*) de Munos [18] pour résoudre de manière ϵ-optimale les problèmes maximin et minimax.

Approximations de V^* La Lipschitz-continuité de V^* permet de dériver des approximateurs à base de points majorant et minorant, à l'aide de cônes pointant respectivement vers le bas et vers le haut. (Imaginer une mâchoire de dents pointues.) Diverses relaxations du zs-POSG à résoudre peuvent alors être envisagées pour dériver les initialisations de ces approximateurs majorant et minorant.

5 HSVI pour zs-POSG

Enfin, la capacité de maintenir de tels approximateurs encadrant (en résolvant les jeux locaux à l'aide d'une optimisation bi-niveau globale) permet de décrire une variante de HSVI pour zs-OMG, donc zs-POSG. Rappelons que HSVI repose sur (i) la génération de trajectoires au cours

1. Nous préférons «jeu de Markov» à «jeu stochastique» à cause du déterminisme de la dynamique.

desquelles chaque joueur agit au mieux d'après l'approximateur qui est optimiste pour lui, et (ii) la mise-à-jour au fur et à mesure les approximateurs jusqu'à atteindre une précision suffisante.
Un critère d'arrêt inspiré par [13] permet de compenser le facteur de branchement infini, et ainsi de démontrer que l'algorithme converge en temps fini vers une solution ϵ-optimale malgré les espaces d'état (d'occupation) et d'action continus.

6 Discussion

Le présent travail démontre théoriquement que l'on peut résoudre des zs-POSG en exploitant la propriété d'optimalité de Bellman. Certains détails d'implémentation comme l'optimisation bi-niveau, l'élagage de cones, ou la possibilité d'utiliser des techniques de compression de l'état d'occupation doivent toutefois être discutés plus avant afin de pouvoir envisager une étude expérimentale de l'approche. En outre, une piste d'amélioration en cours d'étude est la possibilité d'exploiter les propriétés de concavité-convexité de la fonction de valeur optimale démontrées par Wiggers et al. [24] pour obtenir à la fois des approximateurs plus fins et des optimisations bi-niveaux plus efficaces.

Références

[1] N. Basilico, G. De Nittis et N. Gatti : A security game combining patrolling and alarm–triggered responses under spatial and detection uncertainties. Dans *AAAI-16*, 2016.

[2] A. Basu et L. Stettner : Finite- and infinite-horizon Shapley games with nonsymmetric partial observation. *SIAM Journal on Control and Optimization*, 53(6): 3584–3619, 2015.

[3] R. Bellman : On the theory of dynamic programming. *PNAS*, 38:716–719, 1952.

[4] D. Bernstein, R. Givan, N. Immerman et S. Zilberstein : The complexity of decentralized control of Markov decision processes. *Mathematics of Operations Research*, 27(4):819–840, 2002.

[5] N. Brown et T. Sandholm : Superhuman AI for heads-up no-limit poker : Libratus beats top professionals. *Science*, 359(6374):418–424, 2018.

[6] O. Buffet, J. Dibangoye, A. Delage, A. Saffidine et V. Thomas : On Bellman's optimality principle for zs-POSGs, 2020. (à paraître).

[7] O. Buffet, J. Dibangoye, A. Saffidine et V. Thomas : Heuristic search value iteration for zero-sum stochastic games. *IEEE Transactions on Games*, 2020. (à paraître).

[8] K. Chatterjee et L. Doyen : Partial-observation stochastic games : How to win when belief fails. vol. 15, p. 16, 2014.

[9] H. L. Cole et N. Kocherlakota : Dynamic games with hidden actions and hidden states. *Journal of Economic Theory*, 98(1):114–126, 2001.

[10] J. Dibangoye, C. Amato, O. Buffet et F. Charpillet : Optimally solving Dec-POMDPs as continuous-state MDPs. *JAIR*, 55:443–497, 2016.

[11] M. K. Ghosh, D. McDonald et S. Sinha : Zero-sum stochastic games with partial information. *Journal of Optimization Theory and Applications*, 121(1):99–118, avr. 2004.

[12] E. A. Hansen, D. Bernstein et S. Zilberstein : Dynamic programming for partially observable stochastic games. Dans *AAAI-04*, 2004.

[13] K. Horák et B. Bošanský : Solving partially observable stochastic games with public observations. Dans *AAAI-19*, p. 2029–2036, 2019.

[14] K. Horák, B. Bošanský et M. Pěchouček : Heuristic search value iteration for one-sided partially observable stochastic games. Dans *AAAI-17*, p. 558–564, 2017.

[15] H. W. Kuhn : Simplified two-person Poker. Dans H. W. Kuhn et A. W. Tucker, éds : *Contributions to the Theory of Games*, vol. 1. Princeton University Press, 1950.

[16] E. Machuca : An analysis of multiobjective search algorithms and heuristics. Dans *IJCAI-11*, 2011.

[17] O. Madani, S. Hanks et A. Condon : On the undecidability of probabilistic planning and infinite-horizon partially observable Markov decision problems. Dans *AAAI-99*, 1999.

[18] R. Munos : From bandits to Monte-Carlo Tree Search : The optimistic principle applied to optimization and planning. *Foundations and Trends in Machine Learning*, 7(1):1–130, 2014.

[19] K. Åström : Optimal control of Markov processes with incomplete state information. *Journal of Mathematical Analysis and Applications*, 10(1):174 – 205, 1965.

[20] L. S. Shapley : Stochastic games. *PNAS*, 39(10):1095–1100, 1953.

[21] Y. Shoham et K. Leyton-Brown : *Multiagent Systems : Algorithmic, Game-Theoretic, and Logical Foundations*. Cambridge University Press, 2009.

[22] B. S. Stewart et C. C. White, III : Multiobjective A*. *Journal of the ACM*, 38(4):775–814, oct. 1991.

[23] D. Szer, F. Charpillet et S. Zilberstein : MAA* : A heuristic search algorithm for solving decentralized POMDPs. Dans *UAI-05*, p. 576–583, 2005.

[24] A. Wiggers, F. Oliehoek et D. Roijers : Structure in the value function of two-player zero-sum games of incomplete information. Dans *ECAI-16*, p. 1628–1629, 2016.

[25] M. Zinkevich, M. Johanson, M. Bowling et C. Piccione : Regret minimization in games with incomplete information. Dans *NIPS-07*, 2007.